A Study on the Scoring Method of Oral English Test in College English Online Computer Test

Wen Hu(✉)

Hezhou University, Guangxi 542899, China
huwen4431@163.com

Abstract. With the change of English teaching methods, English examination methods have gradually changed from paper examination to computer examination. The online scoring of oral English test is more in line with the needs of teaching reform, and how to score online reasonably is more important. At present, the domestic scoring system technology for Chinese has been quite perfect, and the international scoring method technology for spoken English is not mature. Even if some scoring systems evaluate the scores, there are still some doubts about the accuracy of the scores. In view of the above background and the needs of English proficiency test teaching reform, this paper mainly studies the scoring method of College English online computer-based oral test. At present, most of the scoring methods of spoken English at home and abroad are based on HMM model, which gives a simple machine probability score through the training of corpus, but does not give an objective and specific score. The scoring method of parameter comparison mainly reflects the difference between standard spoken English and standard spoken English, which can not represent the specific score. In view of the above problems, based on the existing achievements and technologies, this paper studies the scoring method of College English online computer-based oral test.

Keywords: College English · Oral test · Network computer test · Scoring method

1 Introduction

Interactive computer-aided language learning system based on speech processing technology is one of the research hotspots of speech technology. It will change the existing language learning environment and teaching mode and greatly improve the efficiency of language learning. The development of the second generation of intelligent voice interaction technology makes our thinking mode no longer single. Whether we can apply voice processing technology to English online examination and realize online English intelligent interactive scoring is the focus of this research. In order to apply the computer-based English test model to college oral English test, it is necessary to study the methods and principles of English scoring [1].

Since the beginning of the pilot work of College English teaching reform of the Ministry of education's College English teaching model based on network and computer in

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2022, LNICST 465, pp. 25–36, 2023.
https://doi.org/10.1007/978-3-031-23950-2_4

February 2004, there has been a revolution in College English teaching and examination in Colleges and universities across the country. In terms of teaching methods, the school has broken the traditional teaching mode and teaching concept [2]. Computer teaching has gone deep into every teaching explanation, vigorously promoted Experiential English, and let students participate in classroom teaching, which fully reflects the interactive teaching concept. In the form of examination, it tends to use the network computer test instead of the paper test, especially in the College English listening and speaking test. It is worth studying whether we can successfully extend the College English online computer test to the College English test. In order to realize the popularization and application of college oral English test online computer test, it is necessary to study how to score the oral English test reasonably and the related principles and algorithms in application [3].

In April 2010, the national "high level Forum on Deepening College English teaching reform and innovating talent training mode" was held in Southeast University. The high-level forum was attended by Liu Xianghong, director of the Liberal Arts Department of the Higher Education Department of the Ministry of education, Jin Guibao, deputy director of the Higher Education Department of the Ministry of education, Professor Jin Yan, director of the CET-4 and CET-6 examination committee, Professor Wang Shouren, director of the college foreign Language Education Steering Committee of the Ministry of education, Professor Yang Zhizhong, President of the National College Foreign Language Teaching Research Association, and Comrade Cao Xiaoying of the College English teaching reform Liaison Office of the Ministry of education, And more than 20 guests from more than 10 universities, including Tsinghua University, Peking University, Nanjing University, Chongqing University, Huazhong University of science and technology, Wuhan University and South China University of technology [4]. Liu Xianghong, director of the liberal arts division of the Department of higher education, from the strategic position of education, the fundamental requirements of education, the strong driving force of education development, the national basic education development policy and the core task of education reform and development, The outline of the national medium and long term education reform and development plan (2010–2020) was publicized in terms of the objectives of the outline of the national medium and long term education reform and development plan. Director Liu Xianghong stressed at the conference to speed up the process of educational informatization and promote the computer teaching and examination mode of College English [5].

In foreign countries, in recent years, more and more investment and achievements have been made in the application research of intelligent voice technology in oral education research. The achievements of English pronunciation evaluation application technology based on the first generation of intelligent voice technology such as MIT and the University of California have been applied in English Autonomous Learning: Several well-known foreign companies include Rosetta stone, aurolog Pearson and others have invested heavily in the research and development of the application of oral English speech testing technology, and some achievements have been industrialized in the market. At the same time, the U.S. government has issued policies to strongly support the R & D and industrialization of English pronunciation evaluation technology, and

has adopted American native technology (phonepass) to test the oral English level of government officials [6].

Based on the educational background described above, this topic is applied under this background. In order to meet the teaching reform of College English grade examination, it will be the final trend to score the examinee's oral English pronunciation by computer. If we can overcome the relevant technical problems, it will have a good market application prospect. Aiming at CET-4 and CET-6, the online computer test is conducted for college students' oral English pronunciation level, and an objective and reasonable evaluation score is given to reflect the candidates' oral English pronunciation level [7]. At present, tablet computers are very popular in the market. If this technology can be embedded into tablet computers to make an application software for scoring oral English pronunciation, which is similar to the form of Chinese learning machine, it will have a good market prospect in China, especially in the field of primary and secondary school students. In addition, there are many promising examples in this list, which shows that the subject is still very promising and worthy of in-depth study.

2 Related work

2.1 Characteristics of Oral English

(1) The emergence and characteristics of oral English

Spoken English refers to the speaker's English pronunciation. Generally speaking, the pronunciation will vibrate. The vibration will compress and expand the air, form sound waves, and spread in the air at a speed of 340 m/s. When the sound wave reaches the human ear, the eardrum will feel the pressure signal of stretching and pressing, and the inner ear nerve will transmit this signal to the human brain, which will recognize the meaning of spoken English. The waveform of oral English pronunciation is irregular for a long time, and the waveform parameters change with time. Therefore, from a macro point of view, oral English pronunciation is a non-stationary state. However, oral pronunciation is generally a certain shape formed by the movement of human oral muscles to form the vocal tract. From this point of view, oral English pronunciation has short-term stability in a certain period of time, which is generally 10 ms–30 ms. For example, the consonants in spoken English can be extracted without regularity by using the characteristics of short-term consonants [8]. The short-term stability of spoken English pronunciation runs through the analysis and processing of spoken English signals. In the process of analyzing and processing spoken English, relevant analysis and research are carried out on this basis. Divide spoken English words into small segments, called frames. The relevant data sampling and feature parameter extraction are carried out for the framing speech.

(2) Acoustic characteristics of spoken English

Spoken English is emitted by the pronunciation organ and has certain physical characteristics, that is, the so-called acoustic characteristics. Aurally speaking, spoken English has the following four characteristics: pitch, length, intensity, and timbre. The sound intensity represents the intensity of the sound. The sound intensity depends on the vibration amplitude of the pronunciation body. The greater the

amplitude, the stronger the pronunciation. In spoken English signal analysis and processing, the energy of speech is expressed by evolving into volume. It can be expressed by the amplitude in a sound frame. Generally, there are two ways to express it as follows:

① The volume is represented by the sum of the absolute values of the sampling points of each sound frame;
② The sum of the squares of the sampling points of each sound frame, then take the logarithm with the base of 10 and multiply by 10.

(3) Pronunciation characteristics of spoken English words.

English: oral pronunciation is a special sound, including sound and language. Spoken English will have a very obvious loud center when speaking, and the voice segments that can be obviously felt are called syllables. Phoneme is a distinguishable basic syllable unit. A syllable of English can contain one or more phonemes. Phonetics uses phonons to represent a concrete realization of phonemes in pronunciation. The phonemes of spoken English words can be divided into vowels and consonants. When the exhaled air flows from the throat, pharyngeal cavity into the mouth and from the lips, these vocal channels are completely open, and the air flow can pass smoothly without obstruction. The phonemes emitted at this time are called vowels. When the air flow is exhaled from the lungs, when a part of the vocal tract is closed, the air flow is blocked, and the air flow cannot pass smoothly. The phonemes that overcome the obstruction of the vocal organs are called consonants. Different phonemes can be distinguished according to the different pronunciation characteristics of vocal tract and vocal cord. The pronunciation characteristics corresponding to these two points include: the state of vocal tract, that is, whether the vocal cord belongs to vibration or opening; The position and height of the tongue, and the specific position of the tongue in the upper jaw, such as the front, middle or rear. Whether the tongue contracts partially or completely. In detail, phonemes are divided into vowels, consonants, diphthongs, affricates and semivowels. If divided according to the type of excitation sound source, the oral English pronunciation generated by periodic glottic wave excitation is called vowel [9]. This kind of oral English signal is driven by vocal cord tension to produce a quasi periodic pulse air flow, which excites the vocal tract. If the vocal cord does not vibrate, but the vocal tract contracts somewhere, forcing the air flow to pass through this contracted part at a higher speed and generating other sound sources, the English pronunciation generated by excitation is called consonant. From the analysis of oral English pronunciation process from a physiological point of view, it can be seen that sound source and vocal tract are the decisive factors of oral English pronunciation.

2.2 Pitch Extraction in Spoken English

Oral English pronunciation is divided into vowels and consonants. The vowel part of English syllable has quasi periodicity, so it can be extracted by pitch. Pitch in spoken English pronunciation is a periodic feature expressed by pronouncing vowels. During oral English pronunciation, the air flow in the lungs passes through the glottis in the

throat, causing periodic vibration of the vocal cords. The periodic pulse train enters the sound channel, and the generated quasi periodic waveform is called pitch. Different pitch reflects different phoneme contents, which can be used as the basis for phoneme identification. On the other hand, rhythm plays an important role in spoken English. English rhythm is based on stress. From the perspective of pronunciation process, the acoustic parameters that play a role in stress perception include at least intensity, duration and pitch change. When comparing the characteristic parameters used at the same time, we also introduce the fundamental frequency trajectory. Therefore, how to extract the pitch of the syllable part of spoken English signal is very important.

Pitch extraction of spoken English signals is mainly divided into time domain method, frequency domain method and time-frequency hybrid method. The time domain method is to estimate the spoken English signal directly to analyze the periodic peak of the waveform, mainly including short-time autocorrelation method, short-time average amplitude difference function method and so on. This paper presents an improved algorithm for pitch extraction of spoken English signals based on autocorrelation (ACF) and average amplitude difference (AMDF). Frequency domain introduces the cepstrum method for pitch extraction. This section will study how to extract the pitch of spoken English signals from time domain and frequency domain.

The pitch extraction method of spoken English signal in time domain focuses on the autocorrelation method (ACF) and average amplitude difference method (AMDF), as well as the improved new algorithm. Based on the principles of ACF and AMDF, the new algorithm makes full use of the clipping idea, and carries out amplitude clipping on the reciprocal of the modified autocorrelation function and the modified average amplitude difference function. The unnecessary extreme value is removed, leaving only the extreme value of the pitch period position. Finally, the extreme value is more prominent through the product of the two to improve the extraction effect.

$$Fn(k) = N + m_1 m_4 \frac{1 - e^{(pk_f + m_2 + m_3 - \lambda)t}}{pk_f + m_2 + m_3 - \lambda} \tag{1}$$

N is the length of the short-time signal. It is easy to see that the autocorrelation function R (k) has a peak at the integer multiple of the pitch period, and the average amplitude difference function f (k) is at the integer multiple of the pitch period A valley value appears on the. Similarly, the modified autocorrelation function RN (k) will have a peak at the integer multiple of the pitch period, while the modified average amplitude difference function fn (k) will have a valley at the integer multiple of the pitch period. If FN (k) is divided by RN (k), the valley will be more obvious. The improved algorithm uses this idea and central clipping to accurately extract the pitch of spoken English signals.

No matter doing short-time autocorrelation analysis or calculating the average amplitude difference, the interference caused by the formant characteristics of the channel will lead to the inconsistency between the peak or valley value and the pitch period, which will bring errors to the detection results In order to solve this problem, two methods can be adopted. The first is filtering. First, let the oral English signal pass through a band-pass filter to remove the influence of most formants; the second is central clipping. The low amplitude part of spoken English signal contains a lot of formant information, while the

high amplitude part contains a lot of pitch information. The influence of formant on the detection results can be effectively reduced by central clipping.

In recent years, an improved pitch extraction algorithm based on cepstrum has been proposed. The idea of improved algorithm can be divided into three categories: 1. The idea and method of statistical detection; 2. The idea of nonlinear channel system model; 3. The method of combining single-sided ACF in cepstrum domain. Figure 1 is a block diagram of an improved pitch extraction algorithm based on cepstrum. This algorithm combines LPC analysis and inverse filtering, and can accurately obtain pitch information.

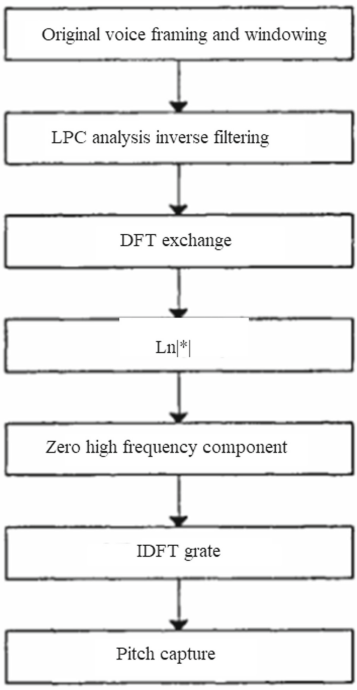

Fig. 1. Pitch extraction based on Cepstrum

3 Research on Spoken English Scoring Algorithm Based on HMM

This paper studies how to score oral English pronunciation. Based on HMM model, two new algorithms are proposed and studied. One of the two new scoring methods is based on phoneme level, the other is based on log likelihood distribution, and the two algorithms are verified by experiments. HMM scoring method is a scoring method based on statistical model. It mainly scores phonemes according to the recognition results of acoustic model, and then scores the whole sentence. Most of the scoring methods based on HMM model are used to score phonemes. At present, the common methods mainly include log likelihood score and log a posteriori probability score. The difference

between this scoring method and the feature comparison scoring method with reference oral English is that this scoring method reflects the reader's oral English pronunciation ability to some extent, not just the feature difference between the standard oral English and the standard oral English.

3.1 Evaluation Algorithm of Spoken English Pronunciation

At present, there are many scoring methods for oral English. The more common ones are: log likelihood scoring, log a posteriori probability scoring, segment classification scoring, segment length scoring, fluency scoring, etc. All these scores are known standard spoken English, and the scores are obtained by comparing the similarity between the tested spoken English and it. The given machine scores are required to be reliable and have a good correlation with the scores of experts. The scores obtained in this way only reflect the readers' oral English pronunciation ability, not the best similarity with the standard oral English pronunciation.

Period duration here refers to the duration of phoneme segments in oral English pronunciation. From the perspective of psychology and linguistics, the dullness of oral English pronunciation will affect the fluency of pronunciation and the unnaturalness of pronunciation. The difference of pronunciation between native speakers and non-native speakers will also affect the duration of the period. In addition, a series of errors such as phoneme insertion, deletion and substitution will also occur when readers pronounce oral English, which will also affect the duration of the period. Therefore, the length of time can also be used to evaluate the quality of pronunciation, especially the fluency and naturalness of oral pronunciation. This requires the standard English corpus to calculate the discrete probability distribution of the duration of the relevant phonemes[10].

3.2 Improved Phoneme Scoring Method Based on HMM

Generally speaking, the evaluation of readers' oral English pronunciation mainly focuses on the characteristics of paragraph and supraparagraph. The proposed scoring method is mainly based on the characteristics of readers' oral pronunciation. The scoring method is based on phonemes and has nothing to do with the content of pronunciation. This text independent scoring method is mainly realized by using the standard oral English pronunciation statistical model and confidence measure algorithm. The statistical model usually adopts the HMM model we are familiar with. The confidence measure algorithm mainly gives the final oral English pronunciation score according to the recognition results of the corpus recognizer, that is, the corresponding score is given according to the recognition results. Generally, we call the score obtained in this way as the machine score, which is usually tested for consistency with the score of experts to reflect the reliability of the machine score.

This section proposes a new phoneme based pronunciation scoring method. The algorithm uses Markov distance to define a new pronunciation score from the phoneme level combined with the log likelihood ratio and speed normalized segment length information given by corpus recognition. In order to score spoken English as a whole and get the phoneme score by weighted average, in order to better understand the algorithm, we first introduce the relevant concepts of HMM model.

(1) Hidden Markov model (HM) physical meaning

Hidden Markov model (HMM) is a statistical model, which is a probability model used to describe the statistical characteristics of a random process. HMM is a double random process. One is the observation sequence that can be obtained through observation, and the other is the transition between the state generators that generate these observation sequences, but the observer cannot see the state transition between these state generators. Therefore, such a double stochastic process is called hidden Markov model. Hidden Markov model "can evolve through Markov chain. The speech signal is generated by the sound source through the sound channel processing. The human voice channel characteristics are divided into a limited number of parts or states with stable characteristics, and the short-time signal generated by each state acting on the sound signal depends on the channel physical parameters or speech probability distribution If the characteristic change of the channel is described by the state transition probability of HMM, it can be expressed as: the probability distribution of the observed value of the generated short-term speech signal is expressed by the generation probability of HMM state, then the HMM model can effectively describe the time-varying speech signal. HMM uses the principles of probability and statistics to successfully characterize the overall time-varying nonstationarity and local short-term stationarity of speech signals. It is widely used in speech recognition and scoring English.

(2) Markov chain

Markov chain is a special case of Markov random process, that is, Markov chain is a Markov process with discrete state and time parameters.

$$\Delta x_{k+1}(t) = \int_0^t e^{(pk_f+m_2+m_3)(t-\tau)}(m_1\Delta u_k(\tau) + pd)d\tau \tag{2}$$

The random observation sequence described by Markov chain reflects the future state probability, which is only related to the current state probability, but has nothing to do with the past state.

(3) Principle of improved phoneme scoring algorithm

The phoneme scoring algorithm first extracts the feature parameters of spoken English. In this section, 39 dimensional MFCC feature parameters are selected. Then, according to the content that readers need to pronounce, the phoneme HMM model is spliced into a forced linear matching network, which is recorded as FA. At the same time, a phoneme cycle recognition network without grammatical model restriction is also generated, which is abbreviated as pl. The output of FA network indicates the matching between the reader's oral pronunciation and the standard spoken language on the phoneme. PL network is the background model needed to construct confidence measure, and its output represents the phoneme recognition results of readers' oral pronunciation. The pronunciation evaluation module is used to comprehensively score the phoneme segmentation, likelihood score and segment length information output by FA and PL networks.

4 Research on Oral English Scoring Algorithm Based on Feature Comparison

4.1 Overview of Feature Comparison Algorithm

If the standard spoken English pronunciation is known, we will use it as the reference standard for scoring. Since there will be one standard pronunciation when listening is played in the oral English computer test, usually the standard spoken English pronunciation, the oral English entered by the reader during follow-up reading will be used as the reference oral English, and the similarity between it and the standard oral English will be compared for scoring basis. This method uses the idea of template matching in speech recognition. The more similar the reader is to the standard spoken pronunciation, the higher the score is given. DTW algorithm or dynamic programming algorithm can be used to calculate the difference of characteristic parameters between them.

In the specific operation, the feature comparison algorithm is mainly used. Compare the acoustic characteristic parameters of the reader's oral English with the standard oral English characteristic parameters, calculate the degree of difference between the reader and the standard oral English, and score the reader's oral English accordingly. The specific flow chart of scoring with standard spoken language is shown in Fig. 2 below, which is mainly divided into three parts: the - part is the extraction of feature parameters, which mainly carries out pitch extraction and MFCC extraction for standard spoken language and reader spoken language. The second part compares the features.

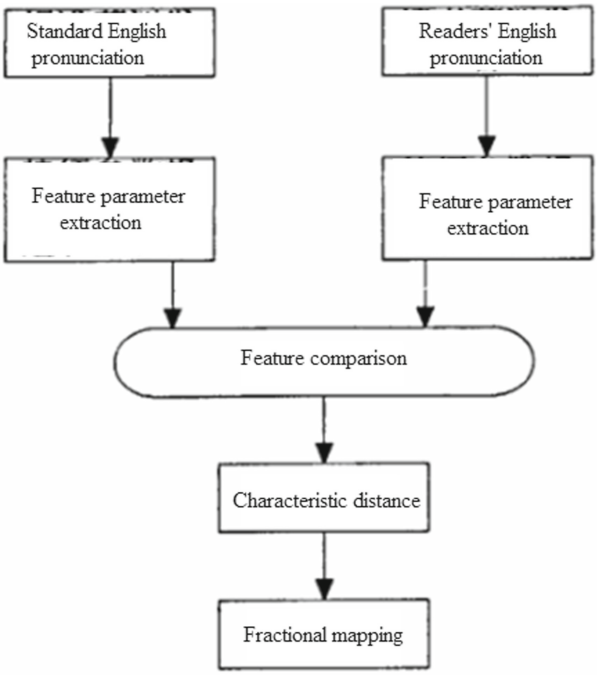

Fig. 2. Block diagram of oral English scoring based on feature comparison

According to the acoustic features extracted above, the common DTW algorithm ‘6 is used to calculate its similarity. Because the feature parameters are extracted by frame, and there is a time length difference between standard spoken language and reader spoken language pronunciation, it is necessary to normalize it, and DTW algorithm is mainly to complete this part of work, carry out template matching and calculate the degree of difference. The third part uses the function mapping method to map the calculated difference distance vector to the objective score, and calculates the correlation coefficient between the score obtained by the algorithm and Mr. Liu of the foreign language department to verify the reliability of the algorithm. The algorithm flow chart is shown in Fig. 2:

4.2 Feature Comparison Algorithm

There are time differences between readers' oral English and standard oral English. What methods should be adopted to avoid the influence of this difference. When extracting the feature parameters of the phoneme level of spoken English signal, the frame number obtained by the same phoneme may only be different between the standard spoken English and the reader's spoken English, or there is a time difference between the standard spoken English and the reader's spoken English. How to reasonably extract and compare the effective frame is very important, The following focuses on the common methods used to deal with such problems, DTW algorithm and its improved algorithm.

(1) Principle of DTW algorithm

Dynamic time warping (DTW) algorithm is a nonlinear optimization method of pattern matching, which successfully solves the problem that the acoustic characteristic parameters vary in the comparison time. The principle of the algorithm is described as follows:

Record the reference template and test template as R = {RL, R2... RM.. RM} and T = {T1, T2..., TN,..., tin} respectively. M represents the total number of frames of the reference template and N represents the total number of frames of the test template. M represents the timing label of the reference frame, and N is the timing label of the test frame. RM represents the m-th frame feature vector of the reference frame, and TN is the n-th frame feature vector of the test frame. Distance is used to reflect the similarity between the two. The smaller the distance, the higher the similarity. Suppose N and m are two frame numbers randomly selected from T and R, respectively. D [Tn, RM] represents the distance between the feature vectors of the two frames. Mark the frame numbers n = 1–N and M = 1–m of the test template and the reference template in a grid. A point (n, m) in the grid represents the intersection of the test template and the reference template frame. Therefore, DTW algorithm is to find the best path from the lower left corner to the upper right corner from the grid. The grid point through which the path passes is the frame number for calculating the distance between the test template and the reference template. As shown in Figure 3

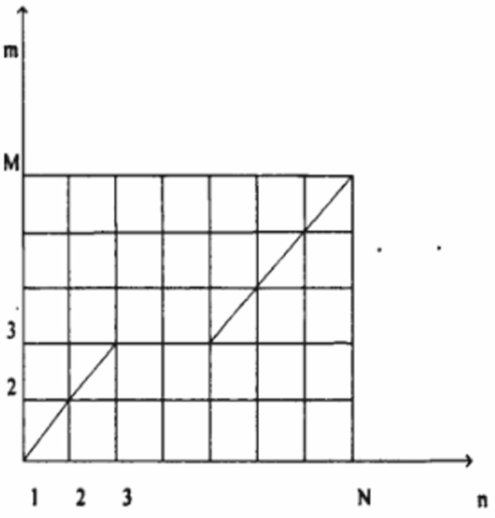

Fig. 3. Schematic diagram of DTW algorithm

The pitch characteristic curve of spoken English intonation is to extract the pitch of the vowel part of the syllable unit. Pitch frequency can be used as all the features of word intonation, but in continuous oral English pronunciation, the intonation of each word syllable unit is affected by the context. The calculation of intonation curve can obtain the pitch frequency of syllable unit, but the conversion information between the front and rear syllables is ignored, that is, the intonation of vowel part does not contain the track characteristics of all intonation. The pitch value of the vowel can be obtained by the pitch frequency of the syllable. Therefore, if the phoneme of the current syllable is a vowel, the frequency between it and the previous syllable is continuous, and the information between the two syllables can be described. If the phoneme of the current syllable is a consonant, the frequency between it and the previous syllable will be separated, which will lose the frequency transfer feature between the two syllable units.

5 Conclusion

Due to the reform of English teaching mode, the way of examination has gradually shifted from paper examination to network computer examination. With the development of the second generation intelligent voice interaction technology, combined with the development needs of domestic English level examination. This paper makes an in-depth study on how to score online in College English online examination. This paper introduces the design of scoring algorithm in detail, and gives some simulation diagram description and data analysis of relevant algorithms, mainly from the two aspects of reference oral English corpus and no reference oral English corpus. Because the oral pronunciation of English is different from that of Chinese, and the components of English words are mainly composed of phonemes. This paper deeply studies the scoring method of spoken English, and puts forward some new algorithms on parameter extraction

and scoring algorithm. However, the research work of this paper mainly focuses on the scoring of some oral English short sentences, and the scoring of an English short passage has not been studied. Because it is only the design of oral scoring algorithm, it is still worth exploring whether it can be well applied to college English online computer-based examination. The diversification of the number of online computer-based examination and the huge voice database how to score online quickly and conveniently are still worth exploring.

Acknowledgements. Horizontalproject of Hezhou University: College English Teaching Reform Innovations in Informatization(2021GX0117) ; Practical Ability Trainings for English Majors of Normal Universities in Professional Accreditation(2019GX0017B) .

References

1. Kido, K., Takahashi, M.: A Study on the Measurement Method of Educational Capability of High School Teachers - ScienceDirect (2021)
2. Zabala-Delgado, J.: A Mixed-Methods Approach to Study the Effects of Rater Training on the Scoring Validity of Local University High-Stakes Writing Tests in Spain (2021)
3. Kim, B.J., Lee, S.B.: A study on the evaluation method of autonomous vehicle for fixed targets (2021)
4. Luo, B.J.: A Study on the innovative design method of rural microlandscape taking edible fungi as an example. E3S Web Conf. **236**(1), 04049 (2021)
5. Park, S.: A study on the method of reflective writing class for korean learners -based on the case of a liberal arts writing subject. Korean Assoc. Gener. Educ. **15**(1), 121–134 (2021)
6. Back, J.D., Jeong, W.M., Lee, J.I., et al.: A Study on the method of calculating the number of working days for harbor construction using the wave of ulsan new port. J. Korean Soc. Coastal Ocean Eng. **33**(2), 80–91 (2021)
7. Hwang, H.J., Khan, J.B., Ji, H.S., et al.: A study on the method of setting the water quality target level for integrated environmental management. J. Korean Soc. Water. Wastewater **35**(3), 187–196 (2021)
8. Cho, S., Lee, J.H.: A study on the usability test method of collaborative robot based on ECG measurement (2021)
9. Hongyan, D., Yangyang, L., Yan, Z., et al.: Correction to: Study on the methodology of striae gravidarum severity evaluation. Biomed. Eng. Online **20**(1), 1–12 (2021)
10. Shi, R.: The role and application of situational teaching in teaching chinese as a foreign language-a case study of primary oral english class. J. Contemp. Educ. Res. **5**(10), 6 (2021)